DOCUMENT RESUME

ED 458 293 TM 033 464

AUTHOR Bishop, N. Scott
TITLE The Validity of Reading Comprehension Test Scores: Evidence of Generalizability across Different Test Administration Conditions.
PUB DATE 2001-04-00
NOTE 36p.; Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Education; *Elementary School Students; Generalizability Theory; *Reading Comprehension; Reading Tests; *Scores; Timed Tests; *Validity

ABSTRACT

This study examined the effects of different test administration conditions on reading comprehension test scores. Evidence of performance differences across district testing conditions might imply that the meanings and interpretations associated with the corresponding test scores have limited generalizability (i.e., knowing how well one reads under one set of conditions might not generalize to performance under other conditions). The issues addressed by this research pertain to the validity of scores from passage-based reading comprehension tests. Three test administration factors were manipulated at grades 3, 5, and 7: (1) whether examinees read a passage before or after the test questions; (2) whether examinees were allowed to review a passage while they were answering questions about it; and (3) whether examinees received prior training (practice) consistent with the conditions under which they took the test. The average number of subjects in each condition was about 66. Iowa Tests of Basic Skills Vocabulary scores were treated as an individual difference variable. The primary dependent variables were total test scores (obtained under standard and extended time limits), skills scores (facts, inferences, and generalizations), and work rates (at 20 and 42 minutes). The results suggest that alternative testing conditions can have complex effects on work rates and test scores that can interact with the ability and grade levels of students. As a main effect, training had little influence on work rates or test scores. No-passage-review administrations were associated with greater working rates, but lower test scores under extended time limits. Questions-first administrations were associated with lower working rates and lower test scores under standard time limits. This finding does not support the beliefs that some hold about the advantages of reading the questions before the passages. (Contains 1 table, 4 figures, and 12 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 293

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

N.S. Bishop

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The Validity of Reading Comprehension Test Scores:

Evidence of Generalizability Across Different Test Administration Conditions

N. Scott Bishop

Riverside Publishing

Paper presented at the Annual Meeting of the

National Council on Measurement in Education

Seattle

April, 2001

TM033464

Requests for reprints may be emailed to: scott_bishop@hmco.com

BEST COPY AVAILABLE

Abstract

This study examined the effects of different test administration conditions on reading comprehension test scores. Evidence of performance differences across distinct testing conditions might imply that the meanings and interpretations associated with the corresponding test scores have limited generalizability (i.e., knowing how well one reads under one set of conditions might not generalize to performance under other conditions). As such, the issues addressed by this research pertain to the validity of scores from passage-based reading comprehension tests. Three test administration factors were manipulated at grades 3, 5, and 7: 1) whether examinees read a test passage before or after the test questions; 2) whether examinees were allowed to review a passage while they were answering questions about it; and 3) whether examinees received prior training (practice) consistent with the conditions under which they took the test. *ITBS* Vocabulary scores were treated as an individual-difference variable. The primary dependent variables were total test scores (obtained under standard and extended time limits), skills scores (facts, inferences, and generalizations), and work rates (at 20 and 42 minutes). The results suggest that alternative testing conditions can have complex affects on work rates and test scores that can interact with the ability and grade levels of students. As a main effect, training had little influence on work rates or test scores. No-passage-review administrations were associated with greater working rates, but lower test scores under extended time limits. Questions-first administrations were associated with lower working rates and lower test scores under standard time limits. This finding does not support the beliefs that some hold about the advantages of reading the questions before the passages.

The Validity of Reading Comprehension Test Scores:

Evidence of Generalizability Across Different Test Administration Conditions

Standard administration procedures for most reading comprehension tests are characterized by two common features: 1) examinees are instructed to read a passage in its entirety before they answer questions about it, and 2) the passage is accessible for the examinees to review while they are answering the questions. Because most reading comprehension tests are intended to be administered in this fashion, an important area of validity research ought to involve determining the generalizability of the test scores obtained from these specific administration conditions to a variety of other reading comprehension contexts.

Test users are probably interested in making inferences from the test scores to a broader domain of skills than are actually represented on the test. What is at issue is the appropriateness of these generalizations. If differential performances were to occur under alternative administration conditions, then the corresponding test scores might not be comparable and the meanings and interpretations associated with them could require qualification. With respect to reading comprehension tests, it may be that inferences involving task demands that are different from those employed under standard testing conditions are limited. Or, as stated in the most recent Standards for Educational and Psychological Testing: "Validation involves careful attention to possible distortions in meaning arising from inadequate representation of the construct and also to aspects of measurement such as test format [and] administration conditions" [italics added] (American Educational Research Association, 1999, p. 9 - 10). The relevance of an investigation into these issues would be bolstered if the alternative administration

conditions studied: 1) were representative of how some examinees actually take these tests (examinee behaviors), and 2) modeled the task demands of typical reading situations.

Examinee Behavior

Regarding examinee behaviors, the format of most reading comprehension tests (with a passage and its associated questions appearing on the same or facing pages in a test booklet) allows examinees some flexibility in how they can actually complete these tests. There is in fact good reason to believe that all examinees do not take passage-based reading comprehension tests in a similar fashion. Farr, Prichard, and Smitten (1990) used think-aloud and retrospective reports to study the test-taking behaviors of 26 college students who took a passage-based reading comprehension test (Iowa Silent Reading Test). Over half of subjects (about 62%) indicated they started taking the test by reading a test passage before answering its associated questions (hereafter called passage first). The most frequent alternative, used by about 27% of the subjects, involved reading the questions before the passage (hereafter called questions first). Some students switched their test-taking behavior during testing. For example, about 31% of the students who started taking the test by reading the passage first abandoned it in favor of a questions-first approach, or a variation similar to it (e.g., reading the questions, then looking at the passage for the answers, one question at a time). No student who started testing reading the questions-first switched from it.

Perhaps of greater importance for large-scale testing programs is the fact that some teachers encourage their students to read the questions before the passages while they are participating in standardized testing. Books written for educators about

classroom literacy assessment (e.g., Anthony, Johnson, Mickelson & Preece, 1991) and test preparation programs (such as Scoring High in Reading from Random House and Taking the (T)error Out of the *ITBS* from American Guidance Service) often advocate that students should read the questions before the test passage (Perlman, Borger, Gonzalez & Junker, 1988 and Perlman, Borger, Gonzalez-Latin, Hiestand, Junker & Rosa, 1989). Although the prevalence of this recommendation at the grade school level is unknown, it may be that entire classrooms, perhaps even schools, are taking reading comprehension tests by reading the questions first (instead of the passages as most directions suggest).

Although it is generally allowed on most standardized tests used at the grade school level, some students may not review the test passages while they are answering test questions. In fact, some studies that have investigated look-back behavior during testing have found that, without training, only a minority of examinees will look back at passages while answering the test questions. For example, Alexander, Hare, and Gardner (1984) found that only about 50% of college students would look back at a test passage while answering questions about it. In another study, the percents of students who referred to the test passages has been as low as 30% and 9%, for good and poor seventh grade readers, respectively (Garner & Reis, 1981). Garner and Hare (1984) and Garner, Hare, Alexander, Haynes, and Winograd (1984) also reported look-back percentages of about 30% for non-trained examinees. These researchers have speculated that the low frequency of look-back behavior during testing occurs because some examinees, particularly younger ones, believe that passage review is not allowed (or would be considered cheating).

Task Demands

Regarding reading task demands, fundamentally different reading situations are encountered both in and outside of school. These situations can differ in important ways, such as the reader's unique goals and purposes for reading, and the actual task demands required of the reader (Wixson, Peters, Weber & Roeber, 1987). Test users should be interested in these issues because the skills of individual readers might vary across different task demands (just as individual reader skills can vary across different reading genres). Current reading comprehension test administration procedures seem to best model situations where readers: 1) process a reading selection before responding to questions about it, and 2) have access to the reading selection while answering questions. (For example, a student completing homework problems from a textbook.) However, in other reading situations, the correspondence to these particular task demands seems weaker. Some reading situations may require comprehension when access to the reading material is not possible (e.g., students taking "closed-book" tests). Also, students occasionally know the questions of interest before reading occurs (e.g., some textbooks give advance questions at the beginning of each chapter). If the skills of individual readers vary in these different contexts, then the scores obtained from testing procedures utilizing a specific combination of task conditions may provide little information about how well one would comprehend under different conditions. In other words, knowing how well one reads under standard testing conditions (passage-first with review allowed) might not generalize to how well one would perform in other situations (such as under questions-first and/or no-review conditions). Correspondingly, if teachers instruct their

students to read the questions-first, the resulting test scores might not be comparable to those from passage-first test administrations.

Previous Research

There has been some limited research about the effects of different administration conditions on reading comprehension test scores, and in particular, on questions-first versus passage-first test administrations. Two hundred and ten (210) students participated in a study conduced by Perlman et al. (1988). Prior to the formal testing, each student received one hour of training and a practice assignment. The ANCOVA results (controlling for prior *ITBS* Reading Comprehension scores) indicated that there was no statistically significant difference between the two groups in their total test performance, or over factual and inferential items. There was, however, an aptitude-by-treatment interaction involving the generalization items. Specifically, students in the passage-first condition with higher pretest scores performed better on the generalization items than higher scoring students in the questions-first condition. The opposite pattern was observed for students with lower pretest scores.

Perlman et al. (1989) investigated three different test-taking conditions. The additional condition arose because the questions-first condition was split into two variants. In one variant, the students only read the item stems. In the other variant, the students read both the stems and the alternatives. Six hundred and six (606) fourth-grade students participated in the study. Prior to testing, each student received 90 minutes of training (divided into two 45-minute sessions) and a practice assignment. The findings were similar to those of the previous study. Specifically, the ANOVA results (using prior *ITBS* Reading Comprehension scores as a blocking factor) indicated that there were no

significant differences among the groups in total test performance or over the factual, inferential, or generalization items.

Students in Grades 3, 5, and 7 participated in the Bishop and Frisbie (1999) study. Unlike the Perlman studies, the Bishop and Frisbie subjects did not receive extensive training and practice (however, the examinees did work through an illustrative sample problem prior to testing.) The testing materials and directions were modified so they would be more congruent with the experimental protocols. For example, the test booklets for the questions-first condition had each question set printed on separate pages that preceded the page on which each passage was contained. Because of the possibility that the testing conditions might differ in efficiency (i.e., the number of test items that examinees can attempt during the allotted testing time), work-rate data was also collected. Substantial differences in favor of the passage-first group were observed at every grade level. For example, the difference in the average total raw scores between the two conditions at grades 3, 5, and 7, were 6.5, 6, and 8, respectively. (The corresponding effect sizes ranged from about 0.4 to 0.9 standard-deviation units.) Differences in working rates and skill scores were of a similar magnitude and also favored the passage-first group.

What has been lacking in the literature is a comprehensive study that manipulates multiple test administration factors. The purpose of the current study is to bridge this gap. This validity study investigated the comparability of reading comprehension test scores (evidence of generalizability across different test administration conditions) by determining what effects different testing conditions have on these test scores at grades 3, 5, and 7. Three test-administration factors were manipulated: 1) whether examinees read

the passage selections before or after the corresponding questions; 2) whether examinees had access to the passage selections while answering questions about them; and 3) whether examinees received prior training (practice) that was consistent with the conditions under which they took the test. *ITBS* Vocabulary scores were treated as an individual-difference variable. The outcomes of interest included:

- Work rates (number of items attempted) at 20- and 42-minutes.
- Total test scores under standard (42-minutes) and extended (62-minutes) test time limits.
- Gain scores (difference between the extended time total test score and standard time total test score).
- Performance on items targeting three different content/process skills (i.e., facts, inferences, and generalizations scores).

The primary research questions included:

1) Do different test administration conditions lead to differences in reading comprehension test performance?
2) Do different test administration conditions lead to differences in working rates?
3) Is the effect of test administration condition on test performance mediated by the grade or ability level of students, or by items targeting different content/process objectives?

Methods

Participants

School systems were selected on the basis of their size, the grades in which testing was conducted, and the school averages from the most recent *ITBS* reading

comprehension test results. Classrooms were then assigned to the experimental conditions so that the expected number of students in each condition (at grades 3, 5, and 7) would be as similar as possible. The average number of subjects in each condition was about 66 (cell sample sizes ranged from a minimum of 40 to a maximum of 75).

Procedures

Subjects at each grade took an on-level version of the *ITBS* Reading Comprehension test (Form H) under one of eight possible test-administration conditions. Test booklets were modified from the standard *ITBS* format so that they would be consistent with the unique administration protocols for each condition. Specifically, the written and oral directions, and the arrangement of the passage/question sets, were adapted to be congruent with the requirements for each condition. As an example, students in the passage-first with review-allowed condition (which modeled how standard directions intend for students to take these tests) were instructed to read a passage first and had access to the passage while answering the questions. The experimental booklets for this condition were formatted so that each test passage appeared alone, on a separate page, prior to its questions. After first reading a passage, the students had to turn to the next page in the booklets in order to see its associated questions. The passage was reprinted for the students on the page facing the questions so that it would be accessible for review. The other experimental conditions differed from the standard test administration procedures in some way. Therefore, the booklets for the alternate conditions were also formatted so as to be consistent with their specific administration protocols (e.g., in questions-first conditions, the booklets had the questions printed alone, on a separate page, prior to the associated passage). Teachers and proctors monitored

student compliance with the testing procedures. Work-rate data was collected by having students circle (on their answer sheets) the item they last answered at 20 and 42 minutes into testing.

Data Analysis

The effects of Grade level, Reading order, Review, and Training on work rates, total test scores, gain scores, and skills scores were evaluated using procedures that allowed exploration of aptitude-by-treatment interaction effects. Specifically, *ITBS* Vocabulary scores (an indicator of word knowledge which serves as a proxy measure of verbal ability) were used as an individual-difference variable in an ANOVA design. (The correlation between Vocabulary and Reading Comprehension scores across all grades was about 0.64, ranging from a low of 0.55 at Grade 3 to a high of 0.73 at Grade 5). The dependent variables were standardized (mean = 0, standard deviation = 1) within each grade. Because of the number of dependent variables analyzed in this study, an alpha level 0.01 was employed in all statistical tests to help preserve the overall Type I error rate.

Results and Interpretation

The ANOVA results for the 20- and 42-minute work rates are documented in Table 1. As expected, the aptitude variable (Vocabulary) was statistically significant (as it was in all subsequent analyses). Regarding the 20-minute work rate, the only main effect that reached statistical significance at the criterion alpha level was Reading order ($F_{1, 1444} = 82.76$, $p < 0.0001$). Students in passage-first conditions (mean = 0.221) attempted more test items by the 20-minute criterion than students in questions-first conditions (mean = -0.229). The corresponding value of Hedges's g effect size, computed by

dividing the marginal mean difference by the square root of the mean square error, was equal to 0.49. Two aptitude-by-treatment interactions were statistically significant. Regarding the Vocabulary-by-Training interaction ($F_{1, 1444} = 14.53$, $p = 0.0001$), Figure 1A indicates that lower ability students who participated in training activities answered more items by the 20-minute criterion than those who did not participate in training. The opposite pattern was observed for higher ability students. (The aptitude-by-treatment interaction figures were plotted by splitting students into four ranked groups based on the quartiles of the sample's *ITBS* Vocabulary scores.) Regarding the Vocabulary-by-Reading interaction ($F_{1, 1444} = 7.28$, $p < 0.007$), Figure 1B indicates that students in passage-first conditions always attempted more items by the 20-minute criterion than students in questions-first conditions. However, the difference in the number of items attempted between the two groups tended to increase as ability level increased.

There were also two significant interactions involving Grade level. Regarding the Grade- by-Training interaction ($F_{2, 1444} = 6.32$, $p = 0.002$), Figure 1C indicates that students in Grades 3 and 5 who participated in training activities attempted more items by the 20-minute criterion than their counterparts. The opposite pattern was observed at Grade 7. Regarding the Grade-by-Reading interaction ($F_{2, 1444} = 8.22$, $p < 0.0003$), Figure 1D indicates that students in passage-first conditions always attempted more items by the 20-minute criterion than their counterparts. However, the difference between the groups was not as large at Grade 7. Such grade level interactions might have been expected as tests at the higher grade levels tend to have relatively longer reading selections and more items per passage than tests at the lower grade levels.

The results for the 42-minute work rates are particularly important because 42-minutes is the standard time limit for the Form H ITBS Reading Comprehension test. Thus, these work rates give an indication of the number of items that might have been completed by examinees during standard testing time. Reading order was significant for the 42-minute work rates ($F_{1, 1414} = 23.25$, $p < 0.0001$). Once again, students in passage-first conditions (mean = 0.133) attempted more test items by the 42-minute criterion than students in the questions-first conditions (mean = -0.116) (Hedges's g = 0.26). The Review main effect was also statistically significant ($F_{1, 1414} = 13.56$, $p = 0.0002$). Students in no-passage-review conditions (mean = 0.102) attempted more test items than students in passage-review conditions (mean = -0.085) (Hedges's g = 0.20). The only significant interaction involved the Training and Review factors ($F_{1, 1444} = 7.69$, $p = 0.006$). Figure 1E indicates that the difference in the number of items attempted by the 42-minute criterion between the passage-review and no-passage-review was greater when students participated in training.

Why did these work rate differences occur? Perhaps questions-first administrations require a greater amount of reading since the test items had to be read twice—once when the items were previewed and then again when the items were being answered). Alternatively, it may simply be that it takes more time to read passages when one is looking for specific information. Passage-review conditions may have had lower work rates because examinees expended additional time reviewing the test passages (compared to their counterparts who were prohibited for engaging in such reviews). Regardless of the reasons underlying the difference, one might speculate that the test performance of students with lower working rates might be influence by the fact that they

did not answer as many test items. The gain score results discussed further below cast additional light on this issue.

Insert Table 1 about here

Insert Figure 1 about here.

The ANOVA results for total test scores under standard (42 minute) and extended (62 minutes) timing conditions, and the resulting gain score difference, are documented in Table 2. The main effect of Reading order was significant for total test scores under standard time conditions ($F_{1, 1414} = 13.37$, $p = 0.0002$). Students in passage-first conditions (mean = 0.086) answered more items correctly than students in questions-first conditions (mean = -0.064) (Hedges's $g = 0.20$). These results are the same as those observed in the Bishop and Frisbie (1999) study, where the passage-first group also outperformed the questions-first group. However, the current effect sizes are somewhat smaller than those reported by Bishop and Frisbie. This finding does not lend support to the beliefs that some hold about the advantages of reading the questions first.

The Vocabulary-by-Grade interaction was significant ($F_{2, 1414} = 6.93$, $p = 0.001$). Figure 2A indicates that the pattern of the conditional means was similar across all grades. However, Grade 5 students at the lower end of the ability distribution scored somewhat lower than their counterparts while the opposite pattern was observed at the upper end of the ability distribution.

Regarding the gain scores, the main effects of Reading ($F_{1, 1414} = 27.96$, $p < 0.0001$) and Review ($F_{1, 1414} = 16.39$, $p < 0.0001$) were both statistically significant

Students in questions-first conditions (mean = 0.142) experienced greater gains with extended testing time than students in passage-first conditions (mean = -0.139) (Hedges's g = 0.29). Students in passage-review conditions (mean = 0.108) experienced greater gains than students in no-passage-review conditions (mean = -0.105) (Hedges's g = 0.22).

The extended time results are important because they give an indication of what the test performances of the groups would have been like independent of the work-rate differences which may have influenced the standard time results. The only main effect that was statistically significant was the Review factor ($F_{1, 1414} = 16.38$, $p < 0.0001$). Students in the passage-review conditions (mean = 0.097) had higher extended time total scores than students in no-passage-review conditions (mean = -0.065) (Hedges's g = 0.22). Once again, the Vocabulary-by-Grade interaction was significant ($F_{2, 1414} = 7.08$, $p = 0.0009$). Figure 2A revealed a similar pattern to that reported above for this interaction. Regarding the significant Grade-by-Review interaction ($F_{2, 1477} = 5.75$, $p = 0.003$), Figure 2C indicates that Grade 5 and 7 students in passage-review conditions had higher extended time total scores than their counterparts. However, at Grade 3, there was virtually no difference between the two groups.

Insert Table 2 about here.

Insert Figure 2 about here.

The ANOVA results for the facts, inferences, and generalizations scores are documented in Table 3. The Reading main effect was statistically significant for facts

($F_{1, 1414} = 7.68$, $p = 0.006$), inferences ($F_{1, 1414} = 11.53$, $p = 0.0007$), and generalizations ($F_{1, 1414} = 15.67$, $p < 0.0001$). Students in passage-first conditions had higher scores than their counterparts over the facts (0.073 vs. –0.046), inferences (0.081 vs. -0.067) and generalizations (0.097 vs. -0.077) items. The corresponding Hedges's g effect sizes were equal to 0.15, 0.19, 0.21, respectively. The Review main effect was only statistically significant for the facts scores ($F_{1, 1414} = 12.18$, $p = 0.0005$). Students in passage-review conditions (mean = 0.088) had higher facts scores than students in no-passage-review conditions (mean = -0.061) (Hedges's $g = 0.18$). One might speculate based on this result that performance on factual items is particularly sensitive to passage review. The *ITBS* interpretive guide notes that the facts items require more literal comprehension on the part of the reader and include processing skills such as understanding factual information and deducing the literal meaning of words from context. Perhaps the difficulty of such tasks increases when the ability to review the corresponding passage is prohibited.

Once again, the Vocabulary-by-Grade interaction was statistically significant for generalizations scores ($F_{2, 1414} = 4.82$, $p = 0.008$) with the same pattern described earlier emerging (see Figure 3A). The Grade-by-Reading ($F_{2, 1414} = 6.74$, $p = 0.001$) and Grade-by-Review ($F_{2, 1414} = 7.19$, $p = 0.001$) interactions were significant for generalizations scores. As Figure 3B indicates, students in passage-first conditions at Grades 3 and 7 had higher generalizations scores than their counterparts, while the difference between the two groups was more similar at Grade 5. Figure 3C indicates that students in passage-review conditions scored higher than their counterparts at Grade 5, but lower than their counterparts at Grade 3. At Grade 7, the performances of the two groups were very

similar. Regarding the significant Training-by-Reading interaction for facts scores ($F_{1,1414} = 7.55$, $p = 0.006$), Figure 3D indicates that students in passage-first conditions who participated in training activities had higher facts scores than their counterparts, while the difference between the two groups was very similar for students who did not participate in training.

Finally, the four-way interaction involving Grade, Training, Reading, and Review reached statistical significance for the facts scores ($F_{2, 1414} = 5.38$, $p = 0.005$). Follow-up ANOVAs conducted at each grade level indicated that the associated triple interaction was only significant at Grade 7 ($F_{1, 486} = 6.84$, $p = 0.009$). Figure 4A indicates that for Grade 7 students who participated in a training session, the passage-first test administrations were associated with higher facts scores than the questions-first administrations. However, for Grade 7 students who did not participate in a training session (Figure 4B), questions-first test administrations were associated with higher facts scores than passage-first test administrations.

Insert Table 3 about here.

Insert Figure 3 about here.

Insert Figure 4 about here.

Discussion

This study has the potential to contribute to our understanding of current assessment practices because the experimental conditions modeled: 1) how some examinees actually take reading comprehension tests, and 2) the task demands of many

everyday-reading situations. Because of the similarity of the *ITBS* reading comprehension test to other reading comprehension tests used in large-scale testing programs, and the range of grade levels employed, this study's findings should generalize broadly to other reading comprehension tests and student populations. For these reasons, many of the parties involved in the testing process (test developers, publishers, and users) are likely to have an interest in the results of this study.

There were statistically significant differences across the test administration conditions studied in terms of work rates, total test scores, gain scores, and skills scores. From this perspective, different administration conditions appear to have complex effects on these outcomes. For example, several aptitude-by-treatment and grade-level interactions were observed. (In fact, if a more liberal significance level had been employed, many other interactions of these types would have been significant.)

From another perspective, some of the observed effect sizes would probably be considered small by commonly cited rules of thumb. They are, however, comparable in magnitude to effect sizes frequently reported in similar research (e.g., studies investigating the effects of coaching, training, and practice on test scores have typically reported effect sizes from about one-tenth to one-third standard-deviation units). Perhaps more importantly, effect sizes should be interpreted relative to other factors, such as the cost of the treatments relative to their effects as well as with respect to the outcome consequences. For example, in a high-stakes testing program, a relatively smaller effect size related to a change in test scores would be more meaningful, particularly if it were a finding replicated across multiple studies.

Overall, it would seem that the differential working rates observed in this study should be an important consideration in test development and score use. Examinee work rates seem sensitive to differences in test administration conditions. This is an important point as the differences in the number of items attempted clearly impacted some of the other study outcomes. When test time limits may not be adequate, performance differences could occur that would call into question the comparability of the test scores.

Regarding this point, it is instructive to further examine the work rate findings in conjunction with the gain score results. Testing conditions that had lower work rates were associated with greater gain scores. For example, the students in questions-first conditions had significantly lower work rates and lower standard-time total test scores than their counterparts, but greater gain scores. Based on their gain scores, and the fact there was no statistically significant difference between these groups with respect to their extended time total test scores, these students clearly benefited from the additional testing time. Similarly, passage-review students also tended to have lower work rates, and as noted above, and experienced greater gain scores. And although not statistically significant, the passage-review students had higher total test scores at 42 minutes than their counterparts. Put simply, the questions-first students, as a group, were able to close the gap in the test scores between themselves and the passage-first students when additional testing time was allowed. However, the passage-review students were able to increase their advantage over the no-passage-review students even further with extended testing time.

The differences in the working rates may have been a contributing factor in some of the skill score differences as well. This is because the skills items are not always

uniformly distributed within the tests at each grade. As an example, many generalization items appear toward the end of the grade 3 and grade 7 tests. Decisions pertaining to these skills might be affected by such difference as lower skills scores could result from differences in working rates operating in conjunction with the disproportional spread of the skills items within the test. As a consequence, one might erroneously conclude that students (individually or as a group) are performing relatively poorly in a given skill area, when in fact, the scores might have been influenced by these other factors.

The training results are noteworthy in light of the fact that many teachers prepare their students for standardized testing with such activities. As a main effect, training had little, if any, impact on work rates or test performance. Although this might seem surprising, there have been other studies that have failed to demonstrate that pre-test training activities have a substantial effect on actual test performance (Mehrens & Kaminski, 1989). However, training interacted with the other test administration factors manipulated in this study as well as with the ability and grade levels of students. Interestingly, these interactions suggested that under some conditions, training might have a negative impact on performance. Perhaps training can influence how students worked (e.g., inducing a more deliberate and slower working approach). Consequently, educators need to further consider the intended and unintended effects of their test preparation activities.

Training was incorporated as a factor in this study because of the differences observed between the Bishop & Frisbie and the Perlman et al. studies. Specifically, it was speculated that the training utilized in the Perlman studies resulted in more similar performances between the questions-first and passage-first groups. However, the results

of the current study do little to support this speculation. It may be that there were important differences between the training activities used in this study and those used in the Perlman studies. For example, the focus and detail of the training in the Perlman studies likely differed from the training activities in the current study, which consisted of a short practice test with only limited emphasis on specific strategies (such as looking for key words, etc.).

Taking all the findings into consideration (the work rate differences, the gain scores resulting from extended testing time, the magnitude of the effect sizes between testing conditions, replication of performance differences observed in previous studies, etc.), there seems to be reason to be concerned about the comparability of the test scores derived under some test administration conditions. As a consequence, the meanings and interpretations associated with reading comprehension test scores might require qualification (e.g., knowing how well one reads under questions-first and/or no-review conditions does not generalize to how well one would perform under passage-first with review conditions, and vice versa). Common inferences based on test scores (such as status and growth) could be affected by the manner in which students take these tests. For example, if a teacher forms reading groups based on last year's test scores, a student's group placement decision could be affected by the conditions under which students were tested. Similarly, impressions about a student's growth might be different depending on the conditions under which the tests were taken from year to year.

The purpose of this study was to determine whether reading comprehension test scores derived under alternative administration conditions are comparable. It must be strongly emphasized that the intent of this study was not to determine how to maximize

reading comprehension test performance. Teachers may feel pressure to intervene, sometimes by inappropriate methods, to achieve this objective (particularly when the test scores are perceived as having high stakes). However, maximizing test scores is not synonymous with maximizing the validity of the scores. Some efforts directed at maximizing student performance might come at the risk of limiting the generalizability of the test scores. If test users intend on using their students' reading comprehension test scores to make inferences to a broader domain of reading skills, then encouraging students to take these tests in such a specific manner risks lowering the validity of any other inferences about the test scores. Perhaps the only reasonable inferences about test scores derived from a questions-first test administration relate to how well students comprehend in situations where questions are given prior to reading. Similar concerns might be raised about the issue of whether a passage review is allowed.

How should the educational measurement community respond to the non-uniform test administration conditions that are occurring in light of the potential score compatibility issues raised above? One issue regards what testing materials (e.g., test interpretation manuals, instructions booklets, etc.) should communicate to teachers about the kind of test-taking advice that is appropriate, or inappropriate, to give to their students. Perhaps the recommended best practice should be for teachers to administer these tests strictly according to standard directions. This recommendation seems to be in harmony with the Standards as well as other guides to professional practice (e.g., the Code of Professional Responsibility in Educational Measurement and the Code of Fair Testing Practices in Education). Furthermore, suggestions that are not in harmony with the standard testing procedures (such as advocating that students read the questions first

instead of the passages) might be counterproductive and have consequences for the future decisions and inferences that are made using the resulting test scores.

A related issue concerns what oral and written test directions are given to students. Because passage review is allowed on most reading comprehension tests, perhaps test directions should specifically address this point. Emphasizing this fact might make a difference, especially for younger test takers who might be confused about this issue. Additionally, perhaps test directions need to explicitly state that the examinees should not read the questions first.

What, if any, time limits should be used on reading comprehension tests? Given the gain scores observed with extended testing time in this study, test developers might want to consider employing more liberal time limits on these tests. However, this might require that the reading comprehension test be administered at a different time in relation to other tests (if it is part of a test battery). Moreover, such a change could create comparability issues with prior tests that used time limits.

An interesting but more involved response would be for test developers establishing separate norms that would be used when alternative-testing conditions occurred. There is precedence for this measurement practice, as separate norms are often provided for mathematics tests when calculators have been allowed. Collecting data for such a norming study (or for establishing a statistical conversion to make the test scores comparable) would be very costly due to the large sample of students that would be required. Although establishing separate norms may not be warranted at this time, it may deserve serious consideration at some point (e.g., if future evidence continues to suggest

that large numbers of teachers are telling their students to take these tests by reading the questions first).

Perhaps reading comprehension tests should contain a mixture of reading conditions, just as they currently contain a mixture of passage types. Reading comprehension tests include a variety of passage types because test users wish to make generalizations about student comprehension over a variety of reading genre. Including a variety of reading task demands might be justified by the same rationale. That is, if it can be determined that test users are similarly interested in how well students comprehend over a variety of different reading conditions, then it would seem reasonable to incorporate these conditions into testing procedures. There may be a number of practical obstacles to overcome when varied conditions are used in a single test administration. However, the gains in validity may well justify such procedures.

Although the inclusion of a variety of task demands might in some sense be an optimal solution to this problem, at the present time, it is not a possibility. It seems logical that if only a specific set of task demands can be utilized during testing, then those demands should be the ones that are most relevant to the test users. It may be that the task demands that are of most relevant to educators are associated with passage-first, passage-review conditions. If this is the case, and only one set of task demands can be used during testing, there would seem to be no good reason for test administrators to permit questions-first conditions if these are of little relevance to test users.

There is opportunity for additional research in this area. For example, an issue not addressed in the present study is whether different testing conditions rank order examinees the same way. A within-subjects design, with subjects taking tests under

multiple conditions, would be one means of addressing this question. Other task variables that might affect reading comprehension test performance also deserve study (e.g., immediate versus delayed testing after reading, the length of the reading selections; the number of test items associated with each passage; and variations in the targeted skills and purposes for reading).

An important need is to identify the nature of reading situations that are considered important in educational settings. For example, educators may not desire a measure of reading comprehension where passage-review is prohibited. Perhaps the important stakeholders in the testing process should be surveyed to determine their opinions on these matters. Bishop & Frisbie (1999) raised several questions that would be relevant for such stakeholders to consider when evaluating these issues. These include: Which set of test-taking conditions best reflects typical reading situations? What type of reading situation does each condition best reflect? Which conditions reflect reading situations that occur most often in educational settings? The answers to these questions might also help narrow the focus of future research aimed at determining the effect of task demand variables on reading comprehension test performance.

It is also important to learn what specific advice teachers are giving their students regarding how they should take reading comprehension tests (and how prevalent the different kinds of advice are). If such research occurs, several related issues should be investigated concurrently. These issues include, among others: Which teacher-suggested strategies are unique to testing and which are promoted more as general reading strategies? Do teachers engage their students in practice or training activities before testing? What conditions characterize the informal reading techniques used by teachers

to assess their students' reading comprehension (and how are these consistent or inconsistent with the test-taking strategies that they recommend to their students)?

Finally, how do variations in task demands affect test performance in other achievement domains and with other types of stimulus materials (e.g., listening comprehension, maps and diagrams, math problem solving, science experiment analysis, etc.)? Using listening comprehension as an example, some of the same issues investigated in the current study might warrant study (e.g., knowledge of questions in advance, training, etc.). Other task conditions relevant to this area might involve audio only versus audio with video presentations and manipulating various characteristics of the speaker.

References

Alexander, P. A., Hare, V. C., & Gardner, R. (1984). The effects of time, access, and question type on response accuracy and frequency of lookbacks in older, proficient readers. Journal of Reading Behavior, 16(2), 119-130.

American Educational Research Association. (1999). Standards for educational and psychological testing. Washington, DC: Author.

Anthony, R. J., Johnson, T. D., Mickelson, N. I., & Preece, A. (1991). Evaluating literacy: A perspective for change. Portsmouth, NH: Heinemann.

Bishop, N. S. & Frisbie, D. A. (1999, April). The effects of different test-taking conditions on reading comprehension test performance. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27(3), 209-226.

Garner, R., & Hare, V. C. (1984). Efficacy of text lookback training for poor comprehenders at two age levels. Journal of Educational Research, 77(6), 376-381.

Garner, R., Hare, V. C., Alexander, P., Haynes, J., & Winograd, P. (1984). Inducing use of a text lookback strategy among unsuccessful readers. American Educational Research Journal, 21(4), 789-798.

Garner, R. & Reis, R. (1981). Monitoring and resolving comprehension obstacles: An investigation of spontaneous text lookbacks among upper-grade good and poor comprehenders. Reading Research Quarterly, 16(4), 569-582.

Hieronymus, A. N., & Hoover, H. D. (1986). Manual for School Administrators—Forms G/H. Chicago: Riverside Publishing Co.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? Educational Measurement Issues and Practices, 8(1), 14-22.

Perlman, C. L. Borger, J., Gonzalez-Latin, C., Hiestand, N., Junker, L., & Rosa, M. (1989). How distracting are the distractors? A comparison of three test-taking strategies. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Perlman, C. L. Borger, J., Gonzalez, C., & Junker, L. (1988). Should they read the questions first? A comparison of two test-taking strategies for elementary students.

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). New directions in statewide reading assessment. The Reading Teacher, 40, 749–754.

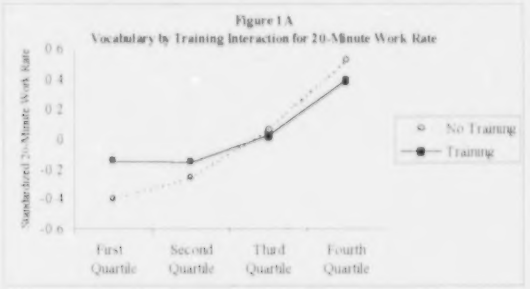
Table 1. ANOVA Results for the 20- and 42-Minute Work Rates.

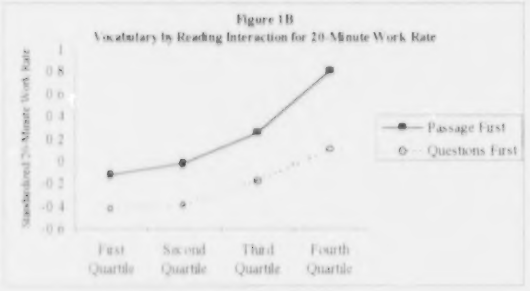
Effect	20-Minute Work Rate	42-Minute Work Rate
Vocabulary (VO)	**$F_{(1,1444)} = 148.24^{a}$**	**$F_{(1,1414)} = 65.4^{a}$**
Grade (GR)	$F_{(2,1444)} = 1.08$	$F_{(2,1414)} = 0.07$
Training (TR)	$F_{(1,1444)} = 1.78$	$F_{(1,1414)} = 0.27$
Reading (RD)	**$F_{(1,1444)} = 82.76^{a}$**	**$F_{(1,1414)} = 23.25^{a}$**
Review (RV)	**$F_{(1,1444)} = 5.12^{d}$**	**$F_{(1,1414)} = 13.56^{b}$**
VO GR	$F_{(2,1444)} = 0.03$	$F_{(2,1414)} = 1.55$
VO TR	**$F_{(1,1444)} = 14.53^{b}$**	**$F_{(1,1414)} = 4.70^{d}$**
VO RD	**$F_{(1,1444)} = 7.28^{c}$**	$F_{(1,1414)} = 1.70$
VO RV	$F_{(1,1444)} = 0.67$	$F_{(1,1414)} = 2.67$
GR TR	**$F_{(2,1444)} = 6.32^{c}$**	$F_{(2,1414)} = 2.20$
GR RD	**$F_{(2,1444)} = 8.22^{b}$**	$F_{(2,1414)} = 1.34$
GR RV	$F_{(2,1444)} = 1.52$	$F_{(2,1414)} = 1.04$
TR RD	**$F_{(1,1444)} = 2.86^{e}$**	$F_{(1,1414)} = 0.72$
TR RV	**$F_{(1,1444)} = 5.20^{d}$**	**$F_{(1,1414)} = 7.69^{c}$**
RD RV	$F_{(1,1444)} = 0.87$	**$F_{(1,1414)} = 2.79^{e}$**
VO GR TR	$F_{(2,1444)} = 1.65$	$F_{(2,1414)} = 1.16$
VO GR RD	$F_{(2,1444)} = 1.35$	**$F_{(2,1414)} = 2.49^{e}$**
VO GR RV	$F_{(2,1444)} = 1.27$	**$F_{(2,1414)} = 2.34^{e}$**
VO TR RD	$F_{(1,1444)} = 0.65$	$F_{(1,1414)} = 1.04$
VO TR RV	**$F_{(1,1444)} = 3.35^{e}$**	**$F_{(1,1414)} = 2.73^{e}$**
VO RD RV	$F_{(1,1444)} = 1.68$	**$F_{(1,1414)} = 4.37^{d}$**
GR TR RD	$F_{(2,1444)} = 2.06$	$F_{(2,1414)} = 0.55$
GR TR RV	$F_{(2,1444)} = 2.19$	$F_{(2,1414)} = 1.15$
GR RD RV	$F_{(2,1444)} = 1.53$	$F_{(2,1414)} = 0.96$
TR RD RV	$F_{(1,1444)} = 0.34$	$F_{(1,1414)} = 0.04$
VO GR TR RD	$F_{(2,1444)} = 1.13$	$F_{(2,1414)} = 1.90$
VO GR TR RV	$F_{(2,1444)} = 1.10$	$F_{(2,1414)} = 0.68$
VO GR RD RV	$F_{(2,1444)} = 0.13$	$F_{(2,1414)} = 0.90$
VO TR RD RV	$F_{(1,1444)} = 2.65$	$F_{(1,1414)} = 0.49$
GR TR RD RV	$F_{(2,1444)} = 0.82$	$F_{(2,1414)} = 1.68$
VO GR TR RD RV	$F_{(2,1444)} = 0.12$	$F_{(2,1414)} = 0.07$

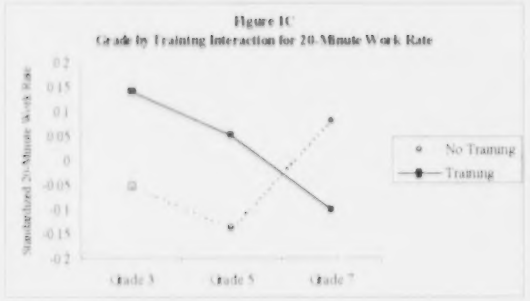
[a] $p < 0.0001$
[b] $p < 0.001$
[c] $p < 0.01$
[d] $p < 0.05$
[e] $p < 0.10$

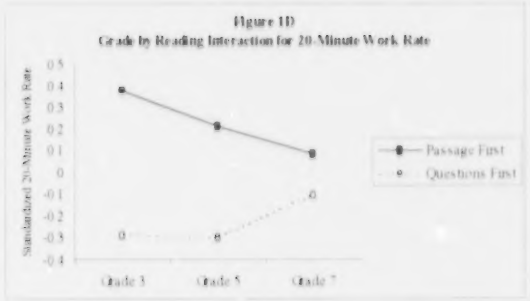
BEST COPY AVAILABLE

Figure 1 Significant 20- and 42-minute work-rate interactions









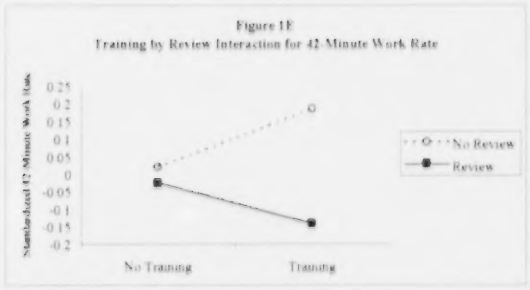


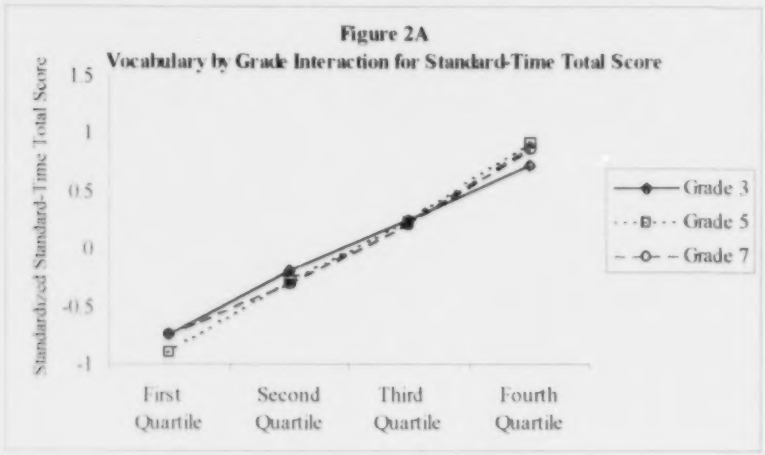
Table 2. ANOVA Results for Standard Time, Extended Time, and Gain Score.

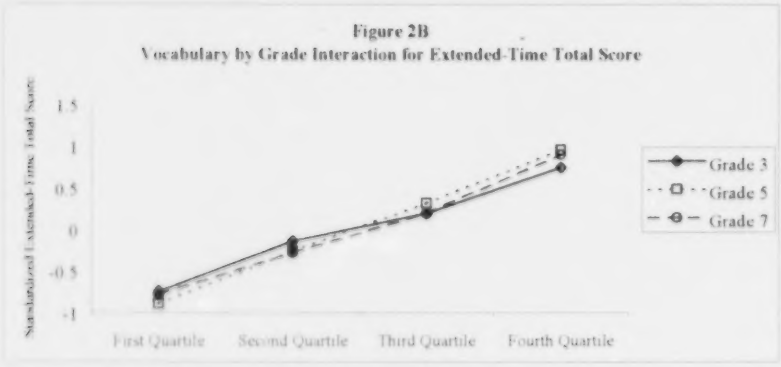
Effect	Standard Time Total Test Score	Gain Score	Extended Time Total Test Score
Vocabulary (VO)	**$F_{(1,1414)} = 932.65^{a}$**	**$F_{(1,1414)} = 13.3^{b}$**	**$F_{(1,1477)} = 1031.03^{a}$**
Grade (GR)	$F_{(2,1414)} = 0.10$	$F_{(2,1414)} = 0.28$	$F_{(2,1477)} = 0.02$
Training (TR)	$F_{(1,1414)} = 0.55$	$F_{(1,1414)} = 0.04$	$F_{(1,1477)} = 0.94$
Reading (RD)	**$F_{(1,1414)} = 13.37^{b}$**	**$F_{(1,1414)} = 27.96^{a}$**	$F_{(1,1477)} = 1.73$
Review (RV)	**$F_{(1,1414)} = 2.72^{c}$**	**$F_{(1,1414)} = 16.39^{b}$**	**$F_{(1,1477)} = 16.38^{b}$**
VO GR	**$F_{(2,1414)} = 6.93^{c}$**	$F_{(2,1414)} = 0.79$	**$F_{(2,1477)} = 7.08^{b}$**
VO TR	$F_{(1,1414)} = 2.63$	$F_{(1,1414)} = 1.09$	$F_{(1,1477)} = 1.39$
VO RD	$F_{(1,1414)} = 0.57$	**$F_{(1,1414)} = 3.91^{d}$**	**$F_{(1,1477)} = 3.47^{e}$**
VO RV	$F_{(1,1414)} = 1.66$	$F_{(1,1414)} = 0.20$	$F_{(1,1477)} = 1.16$
GR TR	$F_{(2,1414)} = 0.09$	$F_{(2,1414)} = 2.12$	$F_{(2,1477)} = 0.07$
GR RD	$F_{(2,1414)} = 2.22$	$F_{(2,1414)} = 1.48$	**$F_{(2,1477)} = 2.42^{e}$**
GR RV	**$F_{(2,1414)} = 4.46^{d}$**	$F_{(2,1414)} = 1.06$	**$F_{(2,1477)} = 5.75^{c}$**
TR RD	**$F_{(1,1414)} = 4.52^{d}$**	$F_{(1,1414)} = 0.04$	**$F_{(1,1477)} = 4.11^{d}$**
TR RV	$F_{(1,1414)} = 0.03$	**$F_{(1,1414)} = 5.30^{d}$**	$F_{(1,1477)} = 0.90$
RD RV	$F_{(1,1414)} = 0.13$	**$F_{(1,1414)} = 2.94^{e}$**	$F_{(1,1477)} = 2.39$
VO GR TR	$F_{(2,1414)} = 0.72$	$F_{(2,1414)} = 1.11$	$F_{(2,1477)} = 0.14$
VO GR RD	$F_{(2,1414)} = 0.83$	$F_{(2,1414)} = 1.85$	$F_{(2,1477)} = 0.27$
VO GR RV	$F_{(2,1414)} = 1.30$	$F_{(2,1414)} = 1.31$	$F_{(2,1477)} = 1.23$
VO TR RD	$F_{(1,1414)} = 0.06$	$F_{(1,1414)} = 1.22$	$F_{(1,1477)} = 0.71$
VO TR RV	$F_{(1,1414)} = 0.01$	**$F_{(1,1414)} = 3.46^{e}$**	$F_{(1,1477)} = 1.21$
VO RD RV	$F_{(1,1414)} = 0.53$	**$F_{(1,1414)} = 3.06^{e}$**	$F_{(1,1477)} = 0.02$
GR TR RD	**$F_{(2,1414)} = 2.73^{e}$**	$F_{(2,1414)} = 0.09$	**$F_{(2,1477)} = 4.11^{d}$**
GR TR RV	$F_{(2,1414)} = 1.39$	$F_{(2,1414)} = 1.93$	**$F_{(2,1477)} = 2.86^{e}$**
GR RD RV	$F_{(2,1414)} = 0.49$	$F_{(2,1414)} = 1.53$	$F_{(2,1477)} = 0.83$
TR RD RV	$F_{(1,1414)} = 0.20$	$F_{(1,1414)} = 0.22$	$F_{(1,1477)} = 0.67$
VO GR TR RD	$F_{(2,1414)} = 0.28$	**$F_{(2,1414)} = 3.18^{d}$**	$F_{(2,1477)} = 0.82$
VO GR TR RV	$F_{(2,1414)} = 1.83$	$F_{(2,1414)} = 1.72$	**$F_{(2,1477)} = 3.14^{d}$**
VO GR RD RV	$F_{(2,1414)} = 1.34$	$F_{(2,1414)} = 0.85$	$F_{(2,1477)} = 1.68$
VO TR RD RV	$F_{(1,1414)} = 0.55$	$F_{(1,1414)} = 1.78$	$F_{(1,1477)} = 0.02$
GR TR RD RV	**$F_{(2,1414)} = 4.24^{d}$**	**$F_{(2,1414)} = 3.62^{d}$**	**$F_{(2,1477)} = 2.34^{e}$**
VO GR TR RD RV	$F_{(2,1414)} = 0.24$	$F_{(2,1414)} = 0.36$	$F_{(2,1477)} = 0.27$

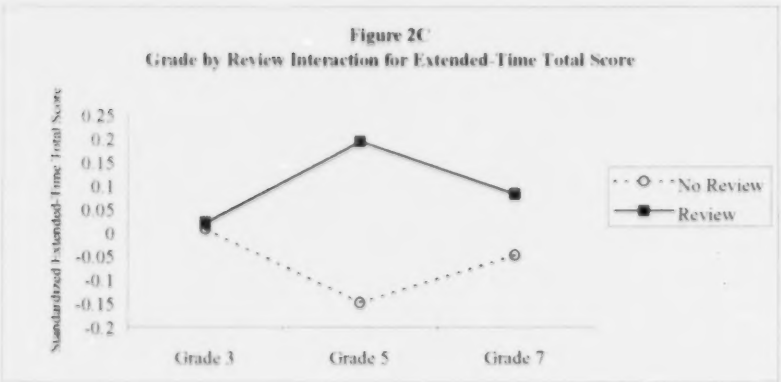
[a] $p < 0.0001$
[b] $p < 0.001$
[c] $p < 0.01$
[d] $p < 0.05$
[e] $p < 0.10$

BEST COPY AVAILABLE

Figure 2. Significant interactions for Standard- and Extended-time Total scores and Gain







BEST COPY AVAILABLE

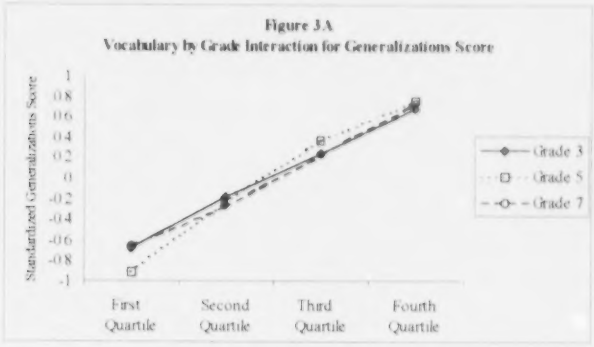
Table 3. ANOVA Results for Skills Scores

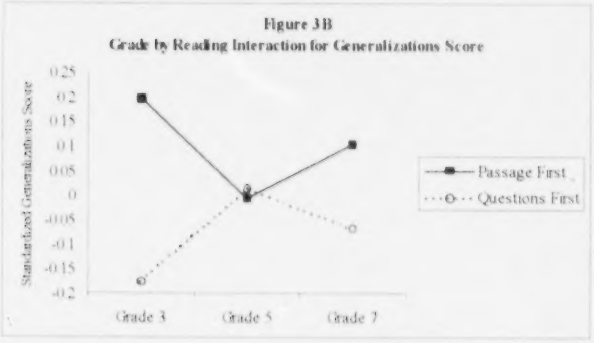
Effect	Facts Score	Inferences Score	Generalizations Score
Vocabulary (VO)	**$F_{(1,1414)} = 786.14$[a]**	**$F_{(1,1414)} = 710.24$[a]**	**$F_{(1,1414)} = 652.25$[a]**
Grade (GR)	$F_{(2,1414)} = 0.29$	$F_{(2,1414)} = 0.15$	$F_{(2,1414)} = 0.04$
Training (TR)	$F_{(1,1414)} = 0.01$	$F_{(1,1414)} = 0.04$	$F_{(1,1414)} = 2.28$
Reading (RD)	**$F_{(1,1414)} = 7.68$[c]**	**$F_{(1,1414)} = 11.53$[b]**	**$F_{(1,1414)} = 15.67$[b]**
Review (RV)	**$F_{(1,1414)} = 12.18$[b]**	$F_{(1,1414)} = 0.89$	$F_{(1,1414)} = 1.43$
VO GR	**$F_{(2,1414)} = 2.95$[e]**	**$F_{(2,1414)} = 3.60$[d]**	**$F_{(2,1414)} = 4.82$[c]**
VO TR	$F_{(1,1414)} = 2.28$	$F_{(1,1414)} = 0.33$	**$F_{(1,1414)} = 4.81$[d]**
VO RD	$F_{(1,1414)} = 0.10$	$F_{(1,1414)} = 0.06$	$F_{(1,1414)} = 0.57$
VO RV	$F_{(1,1414)} = 1.44$	$F_{(1,1414)} = 0.17$	$F_{(1,1414)} = 0.07$
GR TR	$F_{(2,1414)} = 0.28$	$F_{(2,1414)} = 0.33$	$F_{(2,1414)} = 0.36$
GR RD	$F_{(2,1414)} = 1.15$	$F_{(2,1414)} = 0.84$	**$F_{(2,1414)} = 6.74$[c]**
GR RV	**$F_{(2,1414)} = 2.71$[e]**	**$F_{(2,1414)} = 3.16$[d]**	**$F_{(2,1414)} = 7.19$[b]**
TR RD	**$F_{(1,1414)} = 7.55$[c]**	**$F_{(1,1414)} = 4.44$[d]**	**$F_{(1,1414)} = 3.10$[e]**
TR RV	$F_{(1,1414)} = 0.31$	$F_{(1,1414)} = 0.06$	$F_{(1,1414)} = 0.59$
RD RV	$F_{(1,1414)} = 0.11$	$F_{(1,1414)} = 0.54$	$F_{(1,1414)} = 0.92$
VO GR TR	$F_{(2,1414)} = 0.22$	$F_{(2,1414)} = 0.11$	$F_{(2,1414)} = 0.96$
VO GR RD	$F_{(2,1414)} = 1.81$	$F_{(2,1414)} = 0.29$	$F_{(2,1414)} = 0.56$
VO GR RV	$F_{(2,1414)} = 0.21$	$F_{(2,1414)} = 1.57$	$F_{(2,1414)} = 1.24$
VO TR RD	$F_{(1,1414)} = 1.27$	$F_{(1,1414)} = 0.10$	$F_{(1,1414)} = 0.16$
VO TR RV	$F_{(1,1414)} = 0.10$	$F_{(1,1414)} = 1.24$	$F_{(1,1414)} = 0.14$
VO RD RV	$F_{(1,1414)} = 0.15$	$F_{(1,1414)} = 0.00$	$F_{(1,1414)} = 0.50$
GR TR RD	$F_{(2,1414)} = 1.85$	$F_{(2,1414)} = 0.91$	**$F_{(2,1414)} = 3.00$[e]**
GR TR RV	$F_{(2,1414)} = 0.28$	$F_{(2,1414)} = 0.39$	$F_{(2,1414)} = 1.00$
GR RD RV	**$F_{(2,1414)} = 2.53$[e]**	$F_{(2,1414)} = 0.43$	$F_{(2,1414)} = 0.08$
TR RD RV	$F_{(1,1414)} = 0.18$	$F_{(1,1414)} = 0.10$	$F_{(1,1414)} = 0.16$
VO GR TR RD	$F_{(2,1414)} = 1.51$	$F_{(2,1414)} = 0.51$	$F_{(2,1414)} = 0.65$
VO GR TR RV	$F_{(2,1414)} = 1.90$	$F_{(2,1414)} = 1.53$	$F_{(2,1414)} = 0.59$
VO GR RD RV	$F_{(2,1414)} = 0.49$	$F_{(2,1414)} = 1.1$	$F_{(2,1414)} = 0.69$
VO TR RD RV	$F_{(1,1414)} = 2.37$	$F_{(1,1414)} = 0.04$	$F_{(1,1414)} = 0.07$
GR TR RD RV	**$F_{(2,1414)} = 5.38$[c]**	**$F_{(2,1414)} = 4.23$[d]**	$F_{(2,1414)} = 1.67$
VO GR TR RD RV	$F_{(2,1414)} = 0.07$	$F_{(2,1414)} = 0.69$	$F_{(2,1414)} = 0.28$

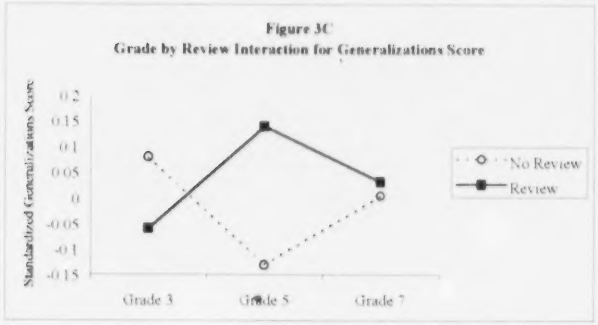
[a] $p < 0.0001$
[b] $p < 0.001$
[c] $p < 0.01$
[d] $p < 0.05$
[e] $p < 0.10$

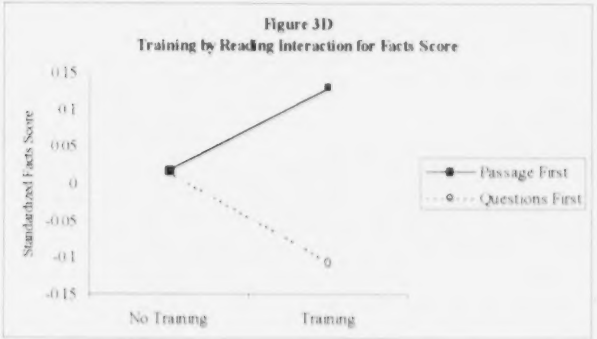
BEST COPY AVAILABLE

Figure 3. Significant Interactions involving Facts, Inferences, and Generalizations Scores



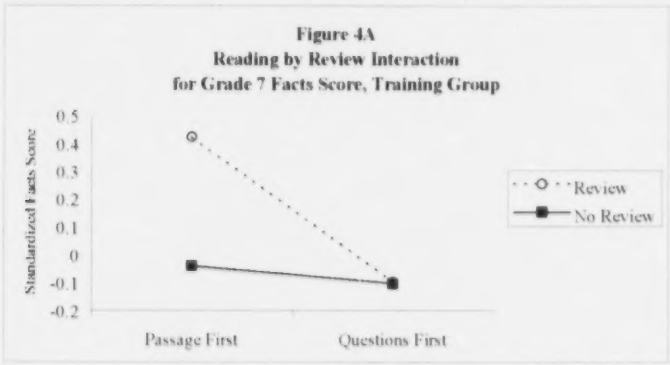


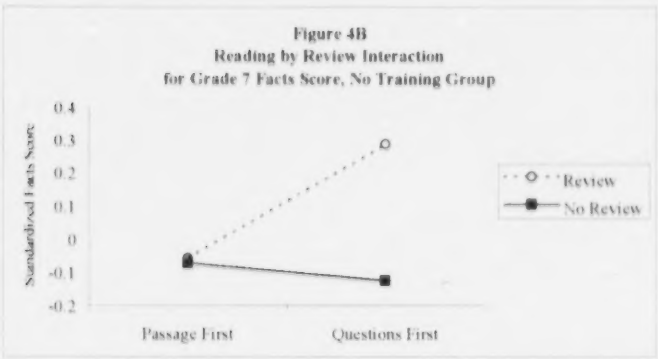




BEST COPY AVAILABLE

Figure 4 Follow-up of the Four-way Interaction for Facts Scores





BEST COPY AVAILABLE